

Chapitre 2

Série statistique à deux variables

(Série Statistique Double)

1. Définition d'une série statistique double	1
2. Quelques exemples	1
3. Représentations de la série double	
3.1 Représentation graphique	2
3.2 Représentation par tableaux de	
3.2.1 Correspondance	2
3.2.2 Contingence	2
3.3.1 Séries Marginales	2
3.3.2 Séries Marginales Conditionnelles	2
3.3.3 Indépendance Stochastique	2
4. Covariance d'une série statistique double	3
5. Ajustement affine	3
5.1 Principe	3
5.2 Méthodes	4
5.2.1 Méthode « au jugé »	4

5.2.2 Méthode de Mayer	4
5.2.3 Méthode des moindres carrés	4
5.2.3.1 Principe	4
5.2.3.2 Détermination des coefficients	5
6. Coefficient de corrélation linéaire	5
7. Ajustement non linéaire	6
7.1 Ajustement Parabolique	6
7.2 Ajustement Logarithmique	6
7.3 Ajustement Exponentiel	6

1. Définition d'une série statistique double

Lorsque l'on s'intéresse à l'étude simultanée de deux caractères d'une même population, on fait ce que l'on appelle des statistiques à deux variables, en étudiant des séries statistiques doubles.

Définition 1 : On considère une population d'effectif n , si on étudie deux caractères X et Y de cette population, on dit que l'on étudie une série statistique double. Chaque individu de cette population est désigné par un nombre compris entre 1 et n . A chaque individu i ($1 \leq i \leq n$) correspond un couple $(x_i ; y_i)$, où x_i est la modalité du caractère X et y_i est la modalité du caractère Y associé à l'individu i .

L'ensemble des couples $(x_i ; y_i)$ définit une série statistique à deux variables.

2. Quelques exemples

Exemple 1 : Le tableau ci-dessous donne, pour chaque ville, le nombre moyen d'heures d'ensoleillement dans l'année, ainsi que la température moyenne :

Ville	Casa	Rabat	Marrakech	Tétouan	Agadir	Nador	Guelmim
Ensoleillement	2072	1729	2763	1574	1685	1833	2790
Température	19.4	18.8	24.2	12.5	13.8	11.7	28.4

- Population : les sept villes étudiées
- Caractère n°1 : nombre moyen d'heures d'ensoleillement dans la ville
- Caractère n°2 : température moyenne dans la ville

Exemple 2 : Le tableau ci-dessous permet de suivre l'évolution de l'espérance de vie à la naissance (en années) au Maroc de 1990 à 1999 pour les femmes:

Année	90	91	92	93	94	95	96	97	98	99
Espérance de Vie	70.1	71.1	71.3	71.4	71.8	71.9	72.0	72.3	72.4	72.4

- Population : les femmes au Maroc
- Caractère n°1 : l'année
- Caractère n°2 : l'espérance de vie

Définition 2 : Lorsque l'un des deux caractères est une année, une date, on dit que la série statistique double est une **série chronologique**.

2. Mise en ordre de la série double

Elle consiste à organiser la suite des couples (x_i, y_i) ; $i=1, \dots, n$ de manière à éviter la redondance. Nous obtenons une nouvelle suite de couples $((x_i', y_j'), n_{ij})$; $i=1, \dots, r$; $j=1, \dots, s$ avec le nombre entier n_{ij} = nombre de couples de la série double qui présentent la valeur x_i' du caractère X et la valeur y_j' du caractère Y .

On a :

$$\sum_{i=1}^r \sum_{j=1}^s n_{ij} = n$$

$X \setminus Y$	y_1		y_j		y_s
x_1	n_{11}	n_{1j}	n_{1s}
x_i	n_{i1}	n_{ij}	n_{is}
x_r	n_{rj}	n_{rj}	n_{rj}

Ce tableau est appelé tableau de correspondance.

Si on remplace les effectifs n_{ij} par les fréquences $f_{ij} = n_{ij}/n$ pour $i=1, \dots, r$ et $j=1, \dots, s$; le tableau obtenu est alors appelé tableau de correspondance.

$$\sum_{i=1}^r \sum_{j=1}^s f_{ij} = 1$$

3. Représentations de la série double

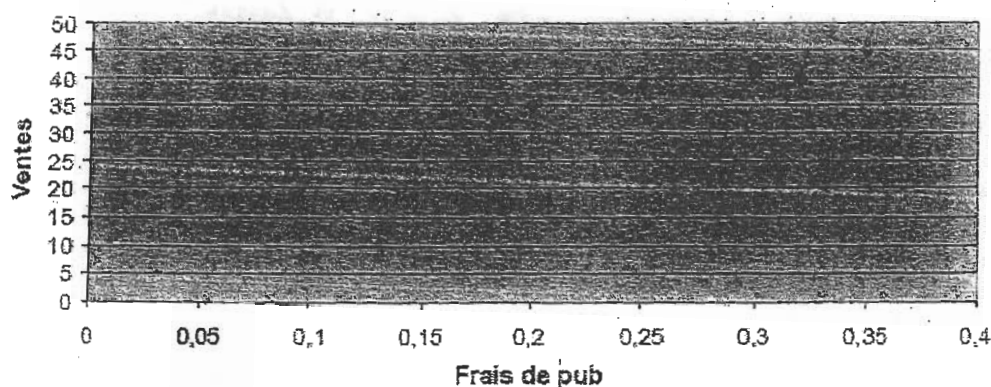
3.1 Représentation graphique

Définition 3 : Si l'on appelle x_1, x_2, \dots, x_n les n valeurs du premier caractère (on notera cette série (x_i)), et si l'on appelle y_1, y_2, \dots, y_n les n valeurs du second caractère (on notera cette série

(y_i)), alors on représente cette série statistique double par un **nuage de points** dans un repère du plan, constitué des points M_i de coordonnées (x_i, y_i)

Exemple 3 : Chaque mois, une entreprise consacre une somme à des opérations publicitaires. On met en regard le montant des ventes chaque mois. Une étude portant sur 8 mois a donné les résultats suivants exprimés en millions de dirhams.

X=Frais de pub	0,24	0,3	0,25	0,32	0,35	0,2	0,18	0,3
Y=Ventes	38	42	39	40	45	35	34	41



Remarque 1

De la donnée de la série statistique double, on peut déduire les séries statistiques simples décrivant séparément les caractères X et Y :

X=Frais de pub	0,18	0,2	0,24	0,25	0,3	0,32	0,35
Effectifs	1	1	1	1	2	1	1

Y=Ventes	34	35	38	39	40	41	42	45
Effectifs	1	1	1	1	1	1	1	1

3.2 Représentation par tableaux

Nous avons deux types de tableaux :

a.- **Tableau de Correspondance** (voir ce qui précède)

b.- **Tableau de Contingence** (voir ce qui précède)

Définition 4 : Soit la moyenne de la série (x_i) , et la moyenne de la série (y_i) . Le point G de coordonnées $(\bar{x} ; \bar{y})$ est appelé **point moyen** du nuage de points associé à cette série statistique double.

3.3.1 Séries Marginales

C'est une série à une dimension où une seule variable varie. Nous en avons 4 types :

3.3.1.1 Série marginale de X (respectivement celle de Y)

Obtenue en faisant la somme des effectifs ou des fréquences des lignes (respectivement des colonnes).

Ainsi l'ensemble des couples $(x_i' ; n_{i.} \text{ ou } f_{i.})$ avec :

$$\sum_{j=1}^s n_{ij} = n_{i.} \text{ et } \sum_{i=1}^r n_{i.} = n \text{ (respectivement } \sum_{j=1}^s f_{ij} = f_{i.} \text{ et } \sum_{i=1}^r f_{i.} = 1)$$

définit la série marginale de X.

De même l'ensemble des couples $(y_j' ; n_{.j} \text{ ou } f_{.j})$ avec :

$$\sum_{i=1}^r n_{ij} = n_{.j} \text{ et } \sum_{j=1}^s n_{.j} = n \text{ (respectivement } \sum_{i=1}^r f_{ij} = f_{.j} \text{ et } \sum_{j=1}^s f_{.j} = 1)$$

définit la série marginale de Y.

3.3.1.2 Série marginale Conditionnelle de X (respectivement celle Conditionnelle de Y)

Obtenue en fixant une colonne du tableau ; par exemple lorsqu'on fixe la valeur de Y à y_j' (colonne n° j).

Ainsi l'ensemble des couples $(x_i' ; f_{i/j})$ avec : $f_{i/j} = \frac{f_{ij}}{f_{.j}}$ avec $\sum_{i=1}^r f_{i/j} = 1$

définit la série marginale Conditionnelle de X lorsque $Y = y_j'$

De même l'ensemble des couples $(y_j' ; f_{j/i})$ avec : $f_{j/i} = \frac{f_{ij}}{f_{i.}}$ avec $\sum_{j=1}^s f_{j/i} = 1$

définit la série marginale Conditionnelle de Y lorsque $X = x_i'$

Remarque

il y a s série marginale de X sachant Y et r série marginale de Y sachant X.

3.3.3 Indépendance Stochastique

Définition 5 : On dit que les caractères X et Y sont indépendants si on a le r*s égalités :

$$\forall i = 1, \dots, r; \forall j = 1, \dots, s; f_{ij} = f_{i.} * f_{.j}$$

Remarque : En pratique, il est plus facile de prouver le contraire !

4. Covariance d'une série statistique double

Définition 6 : On appelle covariance d'une série statistique double (X ; Y) où les caractère X et Y sont quantitatifs le nombre noté cov(X, Y) ou σ_{xy} défini par :

$$\sigma_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Où \bar{x} et \bar{y} sont les moyennes des séries statistiques simples.

Théorème de Huyghens-König

Propriété 1 : Soient $\alpha, \beta, \alpha', \beta'$ des constantes réelles, U et V les caractères statistiques définis par : $U = \alpha X + \beta$ et $V = \alpha' X + \beta'$. C'est-à-dire tels que pour tout i tel que $1 \leq i \leq n$:

$$u_i = \alpha x_i + \beta \text{ et } v_i = \alpha' x_i + \beta'$$

$$\text{Alors : } \text{Cov}(U, V) = \alpha \alpha' \text{Cov}(X, Y)$$

5. Ajustement affine

5.1 Principe

Soit (x_i, y_i) une série statistique double, avec un nuage de points $M_i(x_i, y_i)$ associé.

Lorsque les points du nuage paraissent presque alignés, on peut chercher une relation de la forme $y = ax + b$ qui exprime de façon approchée les valeurs de la série (y_i) en fonction des valeurs de la série (x_i) , autrement dit, une fonction affine f telle que l'égalité $y = f(x)$ s'ajuste au mieux avec les données.

Graphiquement, cela signifie qu'on cherche une droite qui passe au plus près de tous les points du nuage. Une telle relation permettrait notamment de faire des prévisions. Il existe de nombreuses manières d'obtenir un ajustement affine satisfaisant.

5.2 Méthodes

5.2.1 Méthode « au jugé »

A vous de tracer une droite qui passe le plus près possible de tous les points du nuage, si possible en la faisant passer par le point moyen du nuage G . C'est peu précis, mais peut suffire dans certains cas.

5.2.2 Méthode de Mayer

Etape 1 : On commence par « découper » la série statistique double en deux sous-séries bien distinctes, c'est-à-dire que l'on découpe le nuage de points $M_i (x_i, y_i)$ en deux sous-nuages distincts et de même effectif (ou presque : si le nombre de points est pair, pas de souci. S'il est impair, on peut mettre le point surnuméraire dans n'importe lequel des deux sous-nuages).

Etape 2 : On calcule les coordonnées des deux points moyens G_1 et G_2 associés à ces deux sous-nuages, et on place ces deux points sur le graphique.

Etape 3 : On trace la droite $(G_1 G_2)$, appelée droite de Mayer du nuage de points $M_i (x_i, y_i)$, qui doit passer par le point moyen G du nuage de points $M_i (x_i, y_i)$. C'est cette droite qui constitue un ajustement affine tout à fait acceptable pour la série double (x_i, y_i) .

5.2.3 Méthode des moindres carrés

5.2.3.1 Principe

On considère un nuage de points $M_i (x_i, y_i)$ et soit (D) une droite d'équation $y = ax + b$ que l'on cherche à déterminer.

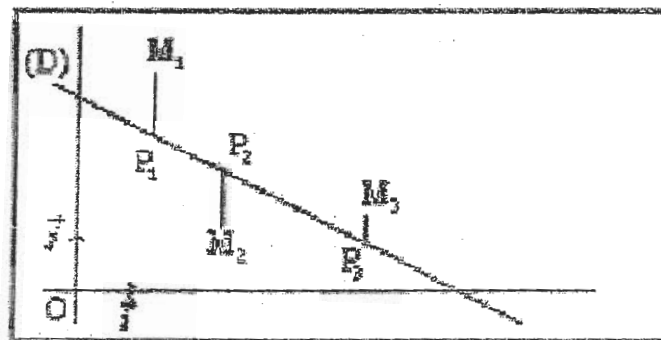
Définition 7 : On appelle somme des résidus associée à la droite (D) , le nombre réel S défini par :

$$S = \sum_{i=1}^n (y_i - (ax_i + b))^2$$

Si P_i désigne le point d'abscisses x_i sur la droite (D) , on a :

$$S = \sum_{i=1}^n M_i p_i^2$$

Définition 8 : On appelle **méthode des moindres carrés** la méthode qui consiste à rechercher les coefficients a et b tels que la somme S soit minimale. Remarquons que S est une fonction des deux variables a et b .



5.2.3.2 Détermination des coefficients

Théorème 1 : Le nombre S est minimum pour

$$a = \frac{\text{cov}(X, Y)}{V(X)} = \frac{\sigma_{XY}}{\sigma_X^2}$$

$$b = \bar{y} - a\bar{x}$$

Proposition 1 : La droite (D) d'équation $y = ax + b$ où a et b sont déterminés d'après théorème 1, est appelé droite de régression de Y en X et on dit qu'on a obtenu cette équation par la méthode des moindres carrés.

Proposition 2 : La droite (D') d'équation : $x = a'y + b'$ avec :

$$a' = \frac{\text{cov}(X, Y)}{V(Y)} = \frac{\sigma_{XY}}{\sigma_Y^2}$$

$$b' = \bar{x} - a'\bar{y}$$

est appelée droite de **droite de régression de X en Y** et on dit qu'on a obtenu cette équation par la méthode des moindres carrés.

Remarque : Les deux droites de régression de Y en X et de X en Y passent toutes deux par le point moyen de coordonnées $(\bar{x} ; \bar{y})$

6. Coefficient de corrélation linéaire

Définition 9 : On appelle coefficient de corrélation linéaire du couple (X, Y) , le nombre réel, noté $r(X, Y)$ tel que :

$$r(x, y) = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y}$$

Remarques

$$-1 \leq r(X, Y) \leq 1$$

$$aa' = (r(X, Y))^2$$

- Lorsque la corrélation est forte ($r^2 \geq 3/4$) les droites de régression sont très proches et le nuage peut être approximé par une droite.
- Lorsque la corrélation est faible, le nuage de points ne peut pas être ajusté par une droite, mais il se peut qu'une autre courbe permette un bon ajustement.

7. Ajustement Parabolique

On considère un nuage de points $M_i (x_i, y_i)$ et soit (P) une parabole d'équation $y = ax^2 + bx + c$ que l'on cherche à déterminer.

Définition 7 : On appelle somme des résidus associée à la parabole (P), le nombre réel S défini par :

$$S = \sum_{i=1}^n (y_i - (ax_i^2 + bx_i + c))^2$$

Si P_i désigne le point d'abscisses x_i sur la parabole (P), on a :

$$S = \sum_{i=1}^n M_i P_i^2$$

Définition 10 : On appelle méthode des moindres carrés la méthode qui consiste à rechercher les coefficients a , b et c tels que la somme S soit minimale. Remarquons que S est une fonction de trois variables a , b et c .

SERIE 2 : Séries Doubles**EXERCICE 1**

Le tableau suivant donne la répartition des salariés d'une entreprise de bâtiment selon le nombre d'enfants à charge (X) et les salaires mensuels perçus (Y) en milliers de dirhams.

Salaires mensuels Y \ Nombre d'enfants X _i	1 - 3	3 - 5	5 - 9
1	4	8	<u>16</u>
2	6	12	24
3	<u>3</u>	6	12
4	2	4	8

- 1 - Donner la distribution marginale de la variable X.
- 2 - Donner la distribution conditionnelle de la variable Y liée à la modalité 4 de X.
- 3 - Que signifient les valeurs 16 et 3 soulignées dans le tableau.
- 4 - Vérifier de deux manières différentes que les deux variables X et Y sont indépendantes.
 - Dites dans ce cas à quoi est égal le coefficient de corrélation linéaire r. (sans le calculer) et dites quelle est la position des deux droites de régression (sans les tracer).
- 5 - Calculer la variance marginale de Y.

Exercice 2

L'observation des prix et des quantités sur un marché de la tomate a donné les résultats suivants :

Quantités x en kg	10	20	35	50	70	90	110	130
Prix y au kg en Dhs	5	3,75	2,75	2,25	1,75	1,25	0,8	0,5

- 1) Les deux caractères X et Y sont-ils linéairement corrélés ? justifier votre réponse.
- 2) Donner la droite de Meyer en prenant G1 point moyen des 4 premiers points de la série double et G2 point moyen des 4 derniers points de la série double. Prévoir le prix d'un kg de tomates pour un achat de 140 kg.
- 3) Déterminer la droite d'ajustement linéaire $y = ax + b$, qui permet d'expliquer le prix au kg par la quantité achetée.
- 4) a. Calculer le coefficient de détermination r^2 et rappeler son signification
 b. Prévoir le prix d'un kg de tomates pour un achat de 140 kg.
 c. Commenter le résultat obtenu.
- 5) Chercher maintenant un ajustement par une fonction logarithme de la forme : $y = a \cdot \ln(x) + b$.
 A.- Calculer a et b par l'une des deux méthodes :
 a.- En posant : $u = \ln(x)$ on se ramènera à un ajustement linéaire :
 $y = a \cdot u + b$.
 b.- Automatiquement par la calculatrice.
 B.- a. Calculer le coefficient de corrélation entre U et Y
 b. Calculer le coefficient de détermination.
 c. Prévoir le prix au kg pour un achat de 140 kg.
- 6) Indiquer lequel de ces trois ajustements vous semble le plus judicieux (on justifiera la réponse).

SERIE 2 : Séries Doubles**EXERCICE 1**

Le tableau suivant donne la répartition des salariés d'une entreprise de bâtiment selon le nombre d'enfants à charge (X) et les salaires mensuels perçus (Y) en milliers de dirhams.

Salaires mensuels Y \ Nombre d'enfants X _i	1 - 3	3 - 5	5 - 9
1	4	8	16
2	6	12	24
3	3	6	12
4	2	4	8

- 1 - Donner la distribution marginale de la variable X.
- 2 - Donner la distribution conditionnelle de la variable Y liée à la modalité 4 de X.
- 3 - Que signifient les valeurs 16 et 3 soulignées dans le tableau.
- 4 - Vérifier de deux manières différentes que les deux variables X et Y sont indépendantes.
 - Dites dans ce cas à quoi est égal le coefficient de corrélation linéaire r. (sans le calculer) et dites quelle est la position des deux droites de régression (sans les tracer).
- 5 - Calculer la variance marginale de Y.

Exercice 2

L'observation des prix et des quantités sur un marché de la tomate a donné les résultats suivants :

Quantités x en kg	10	20	35	50	70	90	110	130
Prix y au kg en Dhs	5	3,75	2,75	2,25	1,75	1,25	0,8	0,5

- 1) Les deux caractères X et Y sont-ils linéairement corrélés ? justifier votre réponse.
- 2) Donner la droite de Meyer en prenant G1 point moyen des 4 premiers points de la série double et G2 point moyen des 4 derniers points de la série double. Prévoir le prix d'un kg de tomates pour un achat de 140 kg.
- 3) Déterminer la droite d'ajustement linéaire $y = ax + b$, qui permet d'expliquer le prix au kg par la quantité achetée.
- 4) a. Calculer le coefficient de détermination r^2 et rappeler son signification
b. Prévoir le prix d'un kg de tomates pour un achat de 140 kg.
c. Commenter le résultat obtenu.
- 5) Chercher maintenant un ajustement par une fonction logarithme de la forme : $y = a \cdot \ln(x) + b$.
A.- Calculer a et b par l'une des deux méthodes :
a.- En posant : $u = \ln(x)$ on se ramènera à un ajustement linéaire : $y = a \cdot u + b$.
b.- Automatiquement par la calculatrice.
B.- a. Calculer le coefficient de corrélation entre U et Y
b. Calculer le coefficient de détermination.
c. Prévoir le prix au kg pour un achat de 140 kg.
- 6) Indiquer lequel de ces trois ajustements vous semble le plus judicieux (on justifiera la réponse).

EXERCICE

L'exploitation fruiticole A. BRICOT doit faire face aux nouvelles réglementations administratives qui permettent l'éligibilité aux aides au secteur primaire.

Dans ce cadre, l'administration vient de demander à A. BRICOT d'établir un relevé statistique des ses arbres fruitiers, reprenant l'âge de l'arbre (en années accomplies) Y et sa hauteur mesurée en centimètres X . Il est à noter qu'aucun arbre de plus de 10 ans et de moins de 3 ans ne peut entrer en ligne de compte pour l'obtention des aides. Les arbres dont il sera question infra sont ceux reconnus par l'administration.

Le formulaire administratif complété se présente comme suit :

Nombre d'arbres de l'exploitation A. BRICOT					
Hauteur (cms)	Age (en années accomplies)				
		[3-4[[5-6[[7-8[[9-10[
	[50 - 100[40	19	3	0
	[100- 150[23	52	8	1
	[150- 200[5	6	24	7
	[200 -300[0	19	6	5

L'administration vous demande :

1. L'âge de la moitié de tous vos arbres. *Que signifie la valeur obtenue ?*
2. La hauteur moyenne de tous vos arbres. *Que signifie la valeur obtenue ?*
3. La classe modale de la hauteur de tous vos arbres. *Que signifie la valeur obtenue ?*
4. La distribution des fréquences de la hauteur de l'arbre conditionnelle à la classe d'âge 7-8 ans.
5. L'âge moyen des arbres conditionnel à une hauteur plus grande que 1 mètre et jusqu'à et y compris 2 mètres.
6. Les variables X et Y sont-elles corrélées ? Justifier votre réponse en déterminant
 - a.- La covariance des deux variables.
 - b.- Le coefficient de corrélation linéaire.
7. La liaison linéaire est-elle justifiée ? Justifier votre réponse.
8. Prévoir la valeur de Y quand X vaut 325 avec le modèle linéaire.